Data Profiling and Data Quality Checks in RStudio

Using R in RStudio

Imported Your Data Already?

- If you already have your data in RStudio, you can skip the slides providing an overview of the import process.

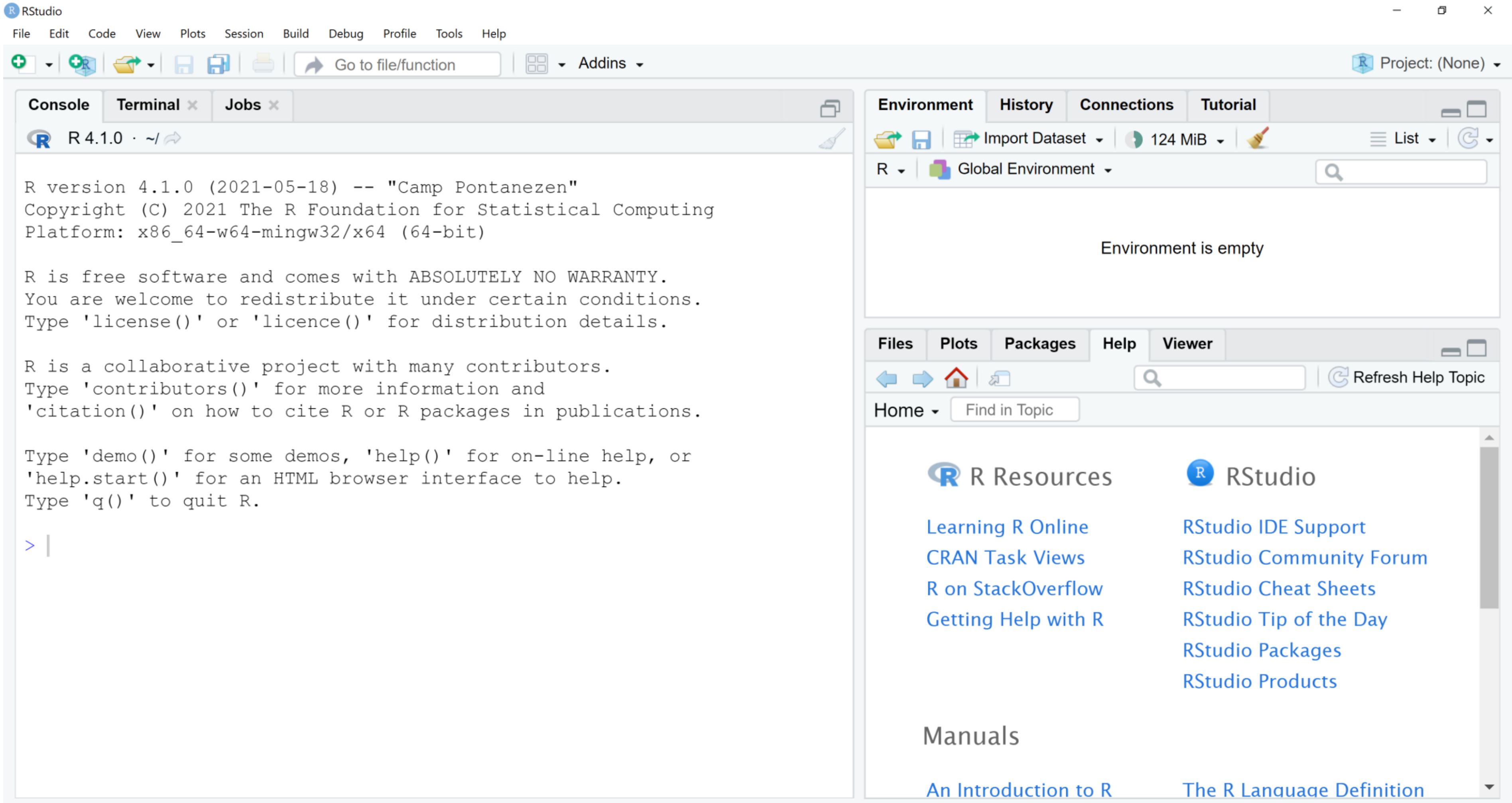
Dataset

- This tutorial is a walkthrough with a sample set of data. You may use this to walk through the tutorial, if you wish, but for your assignments, you will be asked to use your own dataset (as specified within the course).

Dataset reference:

Skoryk, M. (2021). Sepsis Prediction from Clinical Data. Version 1. Retrieved from https://www.kaggle.com/maxskoryk/datasepsis

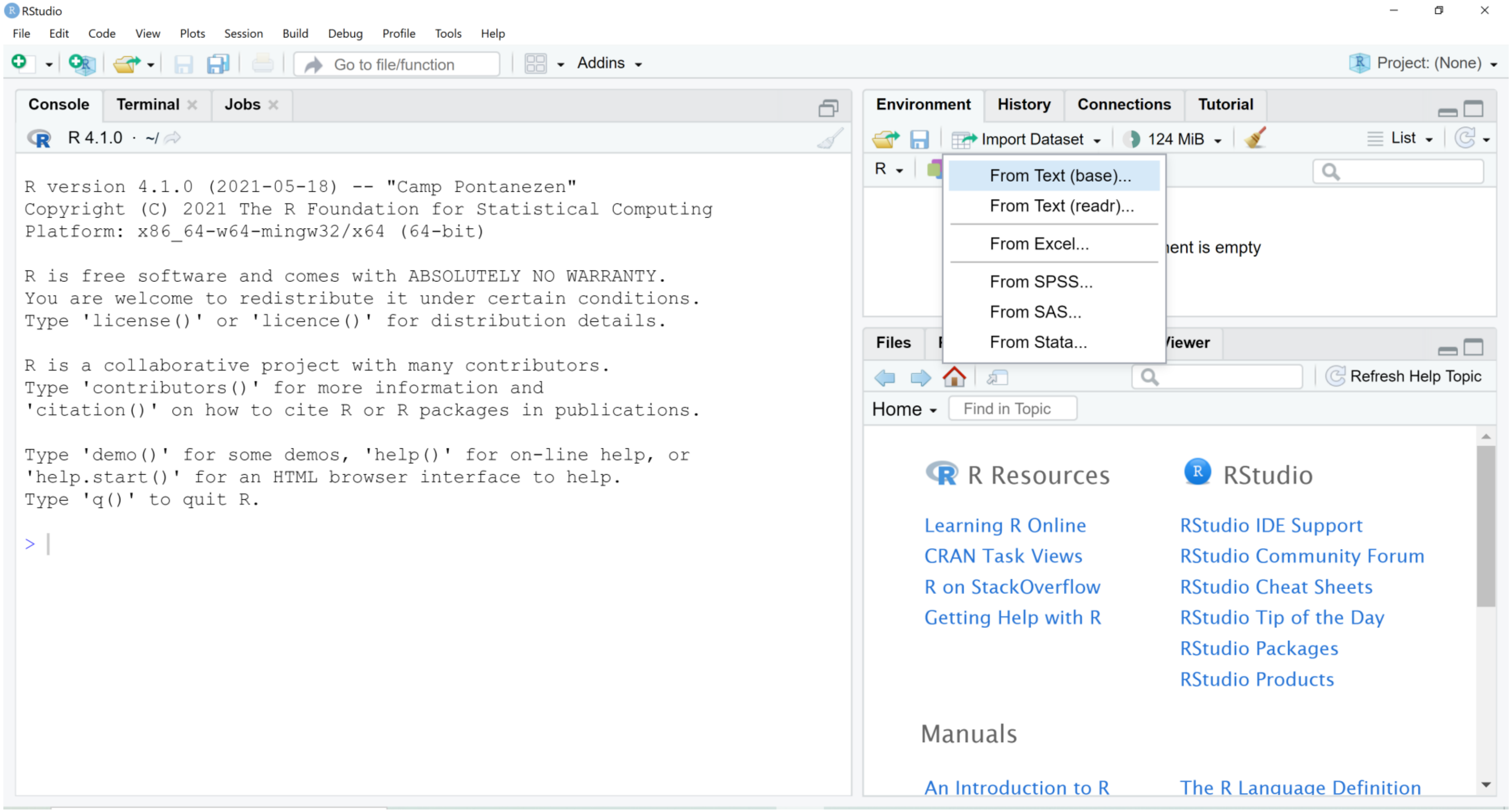
Open R Studio



```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

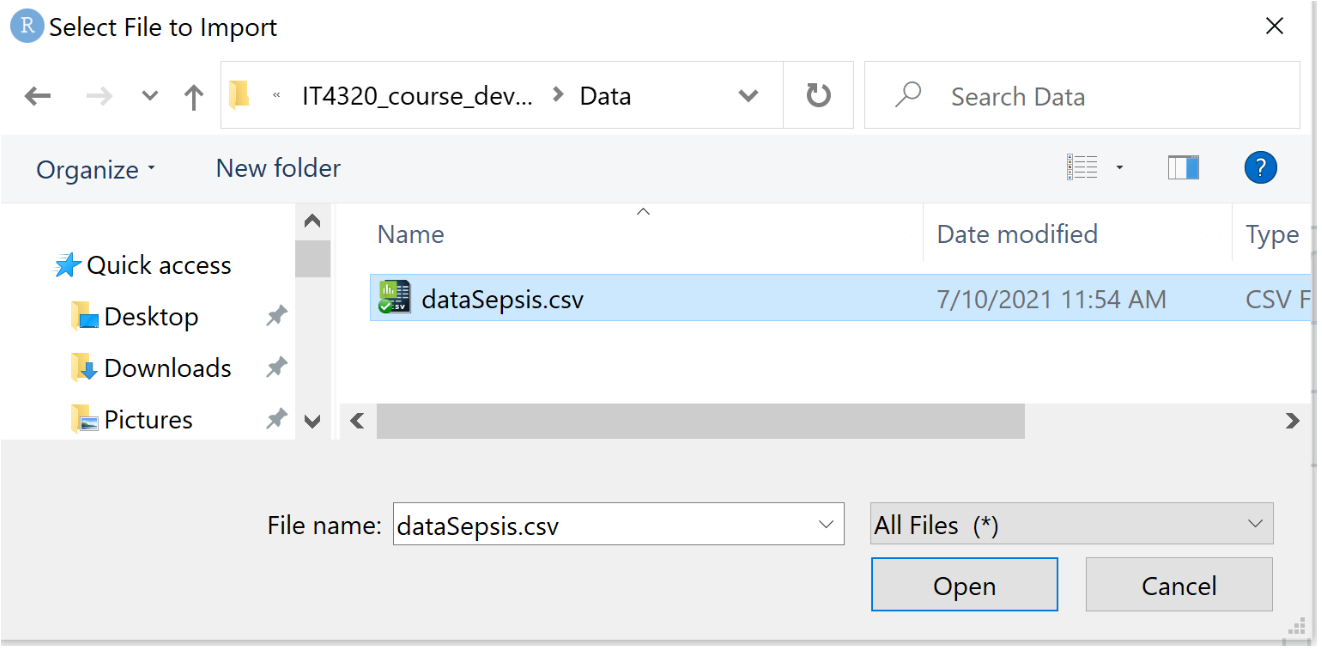
Environment is empty

Click on “Import Data” and Choose “From Text (base)”



```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Navigate to Your Dataset, Then Click “Open”

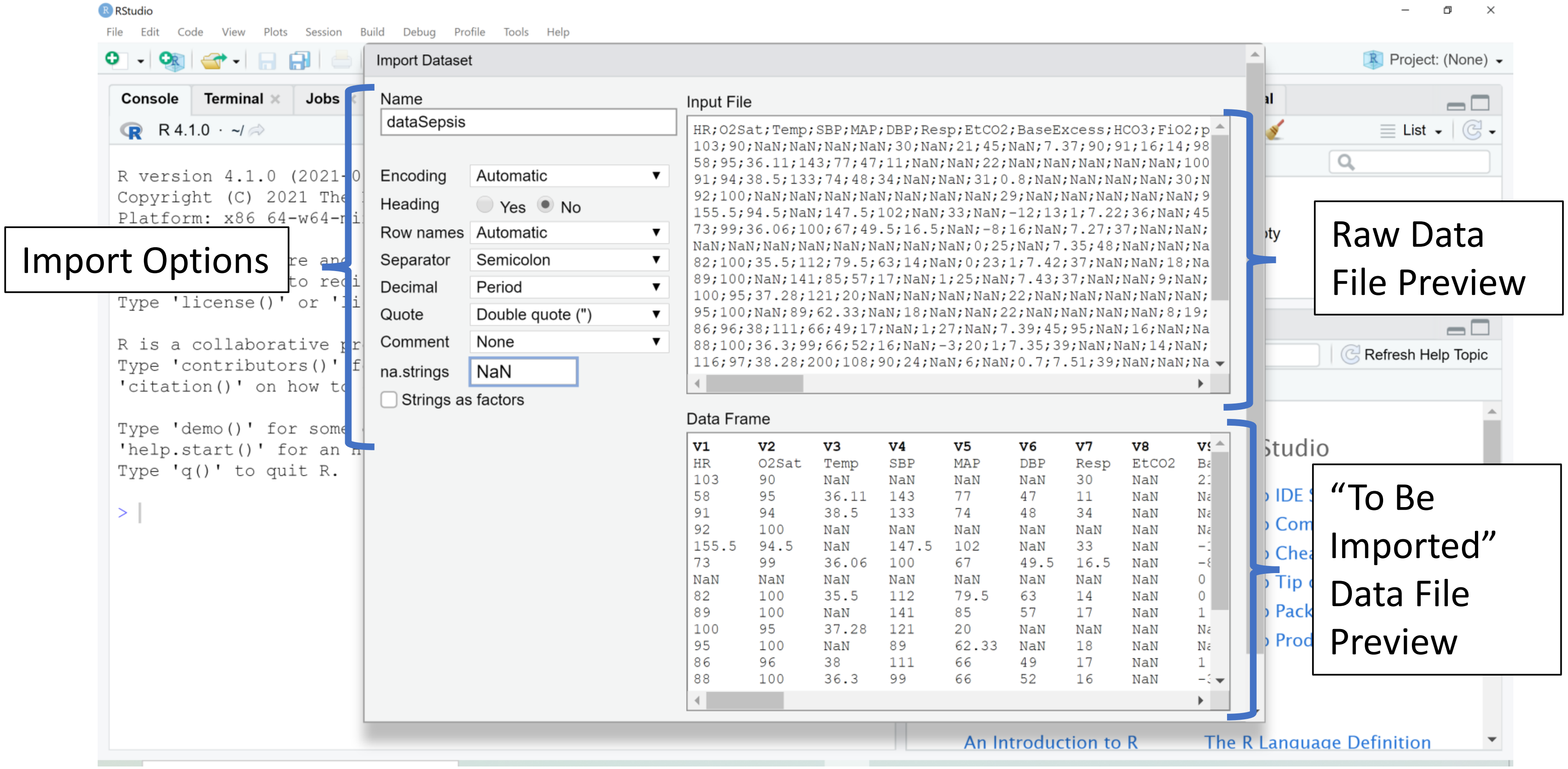


dataSepsis.csv 7/10/2021 11:54 AM CSV F

File name: dataSepsis.csv

All Files (*)

Select Options to Import Your Data Based on the Format of Your Text File



```
R version 4.1.0 (2021-0
Copyright (C) 2021 The
Platform: x86_64-w64-mi
          re and
          to red
Type 'license()' or 'li

R is a collaborative pr
Type 'contributors()' f
'citation()' on how to

Type 'demo()' for some
'help.start()' for an h
Type 'q()' to quit R.

> |
```

Import Dataset

Name: dataSepsis

Encoding: Automatic

Heading: ○ Yes ● No

Row names: Automatic

Separator: Semicolon

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NaN

□ Strings as factors

Input File

```
HR;O2Sat;Temp;SBP;MAP;DBP;Resp;EtCO2;BaseExcess;HCO3;FiO2;p
103;90;NaN;NaN;NaN;NaN;30;NaN;21;45;NaN;7.37;90;91;16;14;98
58;95;36.11;143;77;47;11;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;100
91;94;38.5;133;74;48;34;NaN;NaN;31;0.8;NaN;NaN;NaN;NaN;30;N
92;100;NaN;NaN;NaN;NaN;NaN;NaN;NaN;29;NaN;NaN;NaN;NaN;NaN;9
155.5;94.5;NaN;147.5;102;NaN;33;NaN;-12;13;1;7.22;36;NaN;45
73;99;36.06;100;67;49.5;16.5;NaN;-8;16;NaN;7.27;37;NaN;NaN;
NaN;NaN;NaN;NaN;NaN;NaN;NaN;NaN;0;25;NaN;7.35;48;NaN;NaN;Na
82;100;35.5;112;79.5;63;14;NaN;0;23;1;7.42;37;NaN;NaN;18;Na
89;100;NaN;141;85;57;17;NaN;1;25;NaN;7.43;37;NaN;NaN;9;NaN;
100;95;37.28;121;20;NaN;NaN;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;
95;100;NaN;89;62.33;NaN;18;NaN;NaN;22;NaN;NaN;NaN;NaN;8;19;
86;96;38;111;66;49;17;NaN;1;27;NaN;7.39;45;95;NaN;16;NaN;Na
88;100;36.3;99;66;52;16;NaN;-3;20;1;7.35;39;NaN;NaN;14;NaN;
116;97;38.28;200;108;90;24;NaN;6;NaN;0.7;7.51;39;NaN;NaN;Na
```

Data Frame

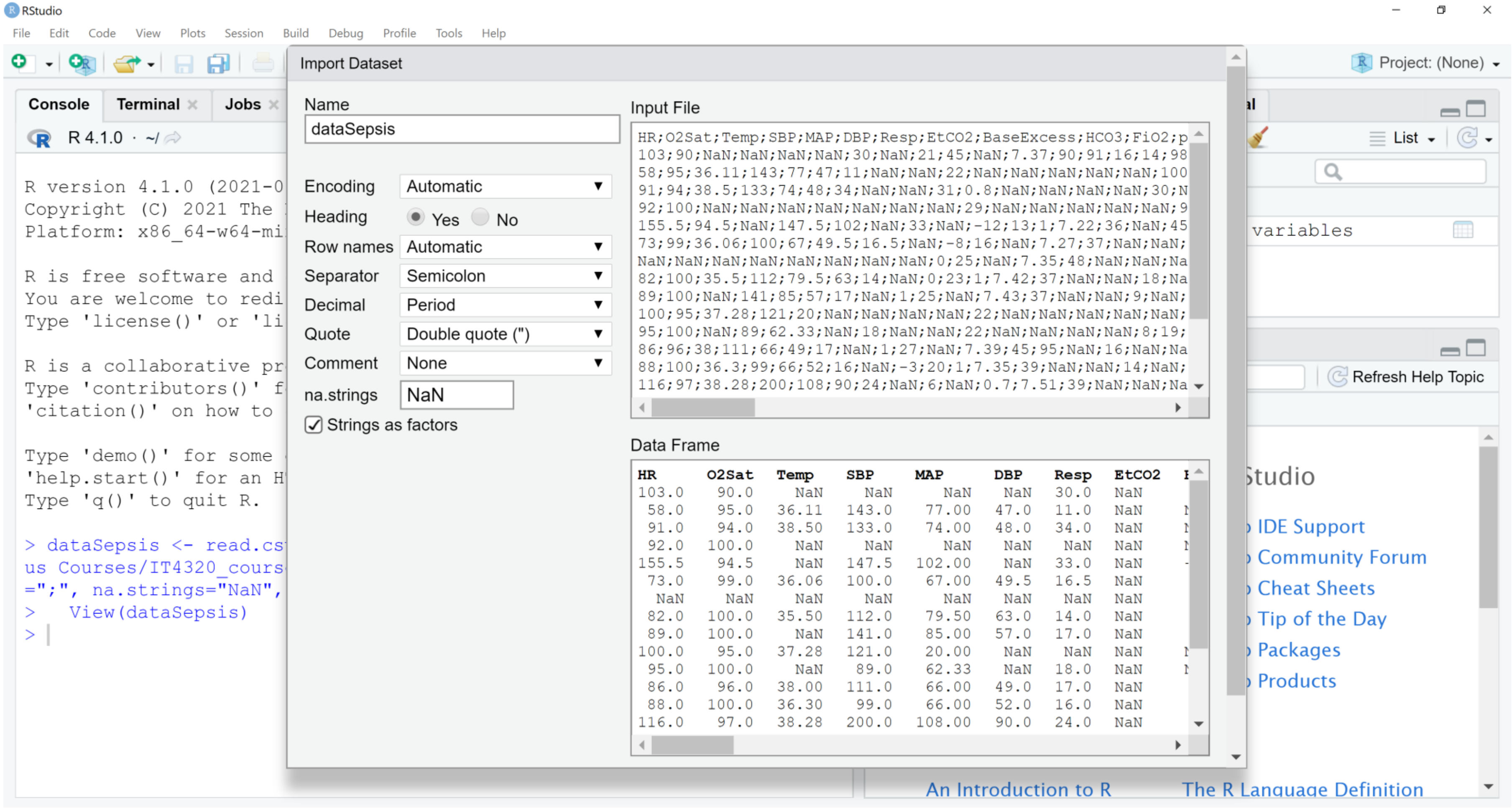
V1	V2	V3	V4	V5	V6	V7	V8	V9
HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	Ba
103	90	NaN	NaN	NaN	NaN	30	NaN	2
58	95	36.11	143	77	47	11	NaN	Na
91	94	38.5	133	74	48	34	NaN	Na
92	100	NaN	NaN	NaN	NaN	NaN	NaN	Na
155.5	94.5	NaN	147.5	102	NaN	33	NaN	-1
73	99	36.06	100	67	49.5	16.5	NaN	-8
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0
82	100	35.5	112	79.5	63	14	NaN	0
89	100	NaN	141	85	57	17	NaN	1
100	95	37.28	121	20	NaN	NaN	NaN	Na
95	100	NaN	89	62.33	NaN	18	NaN	Na
86	96	38	111	66	49	17	NaN	1
88	100	36.3	99	66	52	16	NaN	-3

Import Options

Raw Data File Preview

"To Be Imported" Data File Preview

The Options Displayed are Those Required to Successfully Import the sepsis dataset.



```
R 4.1.0 · ~/

R version 4.1.0 (2021-0
Copyright (C) 2021 The
Platform: x86_64-w64-mi

R is free software and
You are welcome to redi
Type 'license()' or 'li

R is a collaborative pr
Type 'contributors()' f
'citation()' on how to

Type 'demo()' for some
'help.start()' for an H
Type 'q()' to quit R.

> dataSepsis <- read.cs
us Courses/IT4320_cours
=";", na.strings="NaN",
>   View(dataSepsis)
> |
```

Import Dataset

Name: dataSepsis

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Semicolon

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NaN

☑ Strings as factors

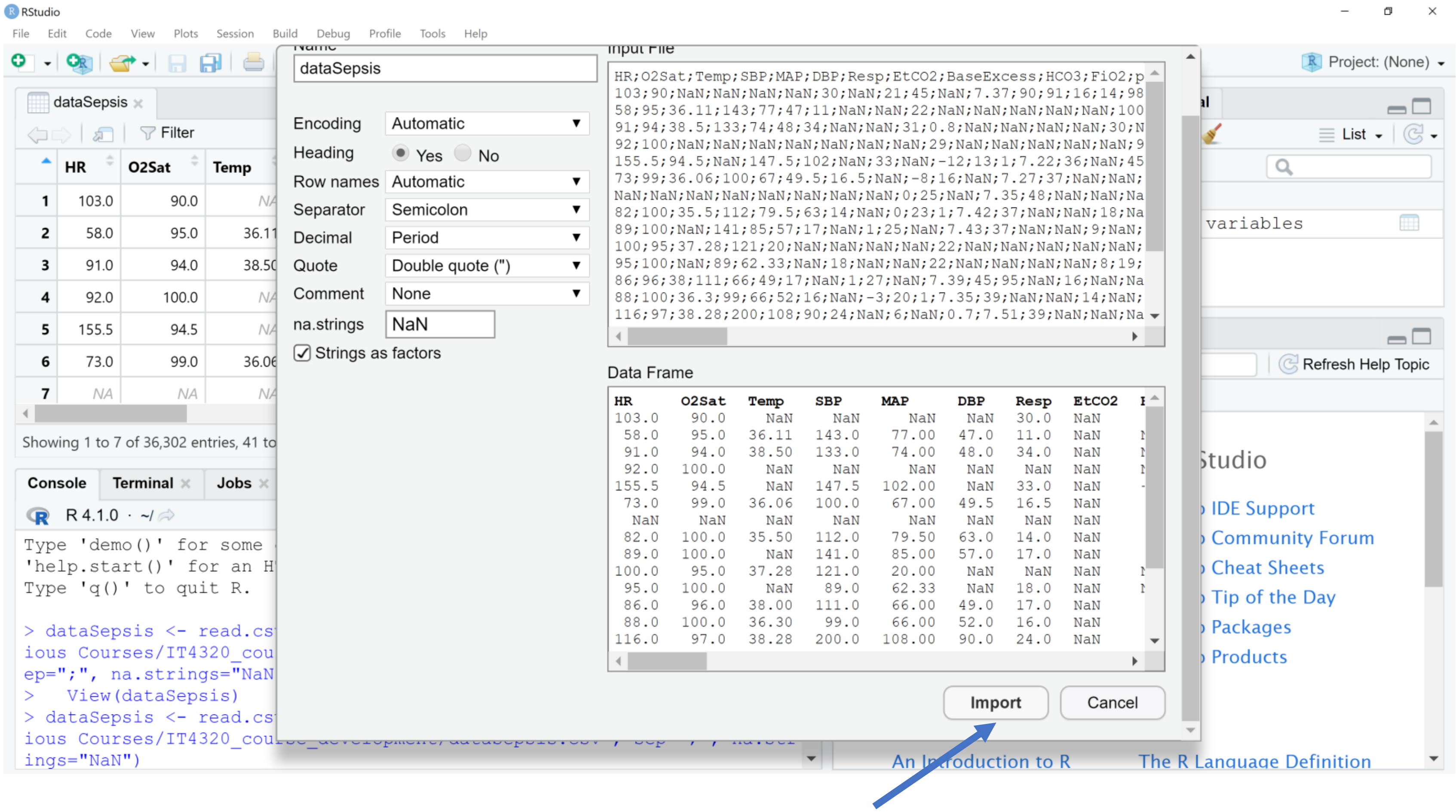
Input File

```
HR;O2Sat;Temp;SBP;MAP;DBP;Resp;EtCO2;BaseExcess;HCO3;FiO2;p
103;90;NaN;NaN;NaN;NaN;30;NaN;21;45;NaN;7.37;90;91;16;14;98
58;95;36.11;143;77;47;11;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;100
91;94;38.5;133;74;48;34;NaN;NaN;31;0.8;NaN;NaN;NaN;NaN;30;N
92;100;NaN;NaN;NaN;NaN;NaN;NaN;NaN;29;NaN;NaN;NaN;NaN;NaN;9
155.5;94.5;NaN;147.5;102;NaN;33;NaN;-12;13;1;7.22;36;NaN;45
73;99;36.06;100;67;49.5;16.5;NaN;-8;16;NaN;7.27;37;NaN;NaN;
NaN;NaN;NaN;NaN;NaN;NaN;NaN;NaN;0;25;NaN;7.35;48;NaN;NaN;Na
82;100;35.5;112;79.5;63;14;NaN;0;23;1;7.42;37;NaN;NaN;18;Na
89;100;NaN;141;85;57;17;NaN;1;25;NaN;7.43;37;NaN;NaN;9;NaN;
100;95;37.28;121;20;NaN;NaN;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;
95;100;NaN;89;62.33;NaN;18;NaN;NaN;22;NaN;NaN;NaN;NaN;8;19;
86;96;38;111;66;49;17;NaN;1;27;NaN;7.39;45;95;NaN;16;NaN;Na
88;100;36.3;99;66;52;16;NaN;-3;20;1;7.35;39;NaN;NaN;14;NaN;
116;97;38.28;200;108;90;24;NaN;6;NaN;0.7;7.51;39;NaN;NaN;Na
```

Data Frame

```
HR      O2Sat   Temp    SBP     MAP      DBP    Resp   EtCO2
103.0    90.0    NaN     NaN      NaN    NaN    30.0   NaN
 58.0    95.0   36.11   143.0    77.00   47.0   11.0   NaN
 91.0    94.0   38.50   133.0    74.00   48.0   34.0   NaN
 92.0   100.0    NaN     NaN      NaN    NaN     NaN   NaN
155.5    94.5    NaN    147.5   102.00   NaN    33.0   NaN
 73.0    99.0   36.06   100.0    67.00   49.5   16.5   NaN
  NaN     NaN    NaN     NaN      NaN    NaN     NaN   NaN
 82.0   100.0   35.50   112.0    79.50   63.0   14.0   NaN
 89.0   100.0    NaN    141.0    85.00   57.0   17.0   NaN
100.0    95.0   37.28   121.0    20.00   NaN     NaN   NaN
 95.0   100.0    NaN     89.0    62.33   NaN    18.0   NaN
 86.0    96.0   38.00   111.0    66.00   49.0   17.0   NaN
 88.0   100.0   36.30    99.0    66.00   52.0   16.0   NaN
116.0    97.0   38.28   200.0   108.00   90.0   24.0   NaN
```

Scroll Down and Click “Import” to Complete Import Process



	HR	O2Sat	Temp
1	103.0	90.0	NA
2	58.0	95.0	36.11
3	91.0	94.0	38.50
4	92.0	100.0	NA
5	155.5	94.5	NA
6	73.0	99.0	36.06
7	NA	NA	NA

Showing 1 to 7 of 36,302 entries, 41 to

```
Type 'demo()' for some
'help.start()' for an H
Type 'q()' to quit R.

> dataSepsis <- read.cs
ious Courses/IT4320_cou
ep=";", na.strings="NaN
>   View(dataSepsis)
> dataSepsis <- read.cs
ious Courses/IT4320_course_development/dataSepsis.csv", sep=";", na.str
ings="NaN")
```

Name: dataSepsis

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Semicolon

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: NaN

☑ Strings as factors

Input File

```
HR;O2Sat;Temp;SBP;MAP;DBP;Resp;EtCO2;BaseExcess;HCO3;FiO2;p
103;90;NaN;NaN;NaN;NaN;30;NaN;21;45;NaN;7.37;90;91;16;14;98
58;95;36.11;143;77;47;11;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;100
91;94;38.5;133;74;48;34;NaN;NaN;31;0.8;NaN;NaN;NaN;NaN;30;N
92;100;NaN;NaN;NaN;NaN;NaN;NaN;NaN;29;NaN;NaN;NaN;NaN;NaN;9
155.5;94.5;NaN;147.5;102;NaN;33;NaN;-12;13;1;7.22;36;NaN;45
73;99;36.06;100;67;49.5;16.5;NaN;-8;16;NaN;7.27;37;NaN;NaN;
NaN;NaN;NaN;NaN;NaN;NaN;NaN;NaN;0;25;NaN;7.35;48;NaN;NaN;Na
82;100;35.5;112;79.5;63;14;NaN;0;23;1;7.42;37;NaN;NaN;18;Na
89;100;NaN;141;85;57;17;NaN;1;25;NaN;7.43;37;NaN;NaN;9;NaN;
100;95;37.28;121;20;NaN;NaN;NaN;NaN;22;NaN;NaN;NaN;NaN;NaN;
95;100;NaN;89;62.33;NaN;18;NaN;NaN;22;NaN;NaN;NaN;NaN;8;19;
86;96;38;111;66;49;17;NaN;1;27;NaN;7.39;45;95;NaN;16;NaN;Na
88;100;36.3;99;66;52;16;NaN;-3;20;1;7.35;39;NaN;NaN;14;NaN;
116;97;38.28;200;108;90;24;NaN;6;NaN;0.7;7.51;39;NaN;NaN;Na
```

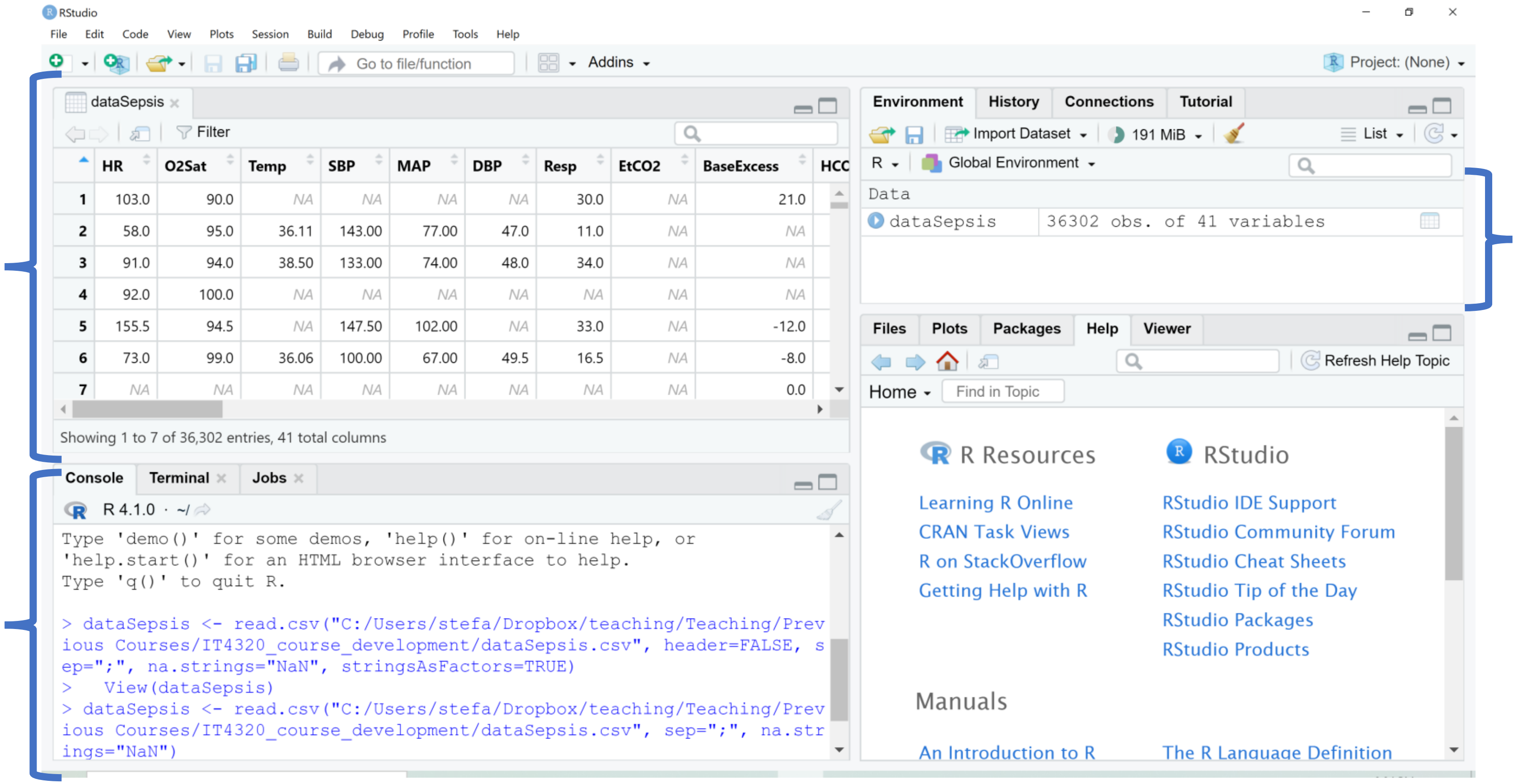
Data Frame

```
HR      O2Sat   Temp    SBP     MAP     DBP    Resp   EtCO2   I
103.0    90.0    NaN     NaN     NaN    NaN    30.0   NaN
 58.0    95.0   36.11   143.0    77.00  47.0   11.0   NaN     N
 91.0    94.0   38.50   133.0    74.00  48.0   34.0   NaN     N
 92.0   100.0    NaN     NaN     NaN    NaN    NaN    NaN     N
155.5    94.5    NaN    147.5   102.00  NaN    33.0   NaN     -
 73.0    99.0   36.06   100.0    67.00  49.5   16.5   NaN
  NaN     NaN    NaN     NaN     NaN    NaN    NaN    NaN
 82.0   100.0   35.50   112.0    79.50  63.0   14.0   NaN
 89.0   100.0    NaN    141.0    85.00  57.0   17.0   NaN
100.0    95.0   37.28   121.0    20.00  NaN    NaN    NaN     N
 95.0   100.0    NaN     89.0    62.33  NaN    18.0   NaN     N
 86.0    96.0   38.00   111.0    66.00  49.0   17.0   NaN
 88.0   100.0   36.30    99.0    66.00  52.0   16.0   NaN
116.0    97.0   38.28   200.0   108.00  90.0   24.0   NaN
```

Import   Cancel

© Stefanie G. Reay, MS, PhD, Capella University, 2021

You May Verify Successful Upload On the Following Screen



	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCC
1	103.0	90.0	NA	NA	NA	NA	30.0	NA	21.0	
2	58.0	95.0	36.11	143.00	77.00	47.0	11.0	NA	NA	
3	91.0	94.0	38.50	133.00	74.00	48.0	34.0	NA	NA	
4	92.0	100.0	NA	NA	NA	NA	NA	NA	NA	
5	155.5	94.5	NA	147.50	102.00	NA	33.0	NA	-12.0	
6	73.0	99.0	36.06	100.00	67.00	49.5	16.5	NA	-8.0	
7	NA	NA	NA	NA	NA	NA	NA	NA	0.0	

Showing 1 to 7 of 36,302 entries, 41 total columns

```
Data
dataSepsis      36302 obs. of 41 variables
```

```
R 4.1.0 · ~/
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> dataSepsis <- read.csv("C:/Users/stefa/Dropbox/teaching/Teaching/Prev
ious Courses/IT4320_course_development/dataSepsis.csv", header=FALSE, s
ep=";", na.strings="NaN", stringsAsFactors=TRUE)
>   View(dataSepsis)
> dataSepsis <- read.csv("C:/Users/stefa/Dropbox/teaching/Teaching/Prev
ious Courses/IT4320_course_development/dataSepsis.csv", sep=";", na.str
ings="NaN")
```

Imported Data

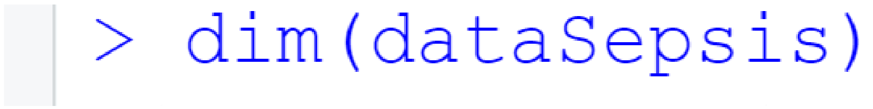
Dataset Summary

Logfile Indicating Options in the Import Process

More Options for Importing Data Into R Studio

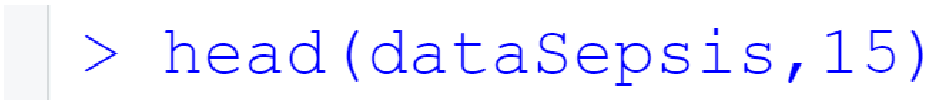
https://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-the-RStudio-IDE

First: Look at the contents of the dataset



```
> dim(dataSepsis)
```

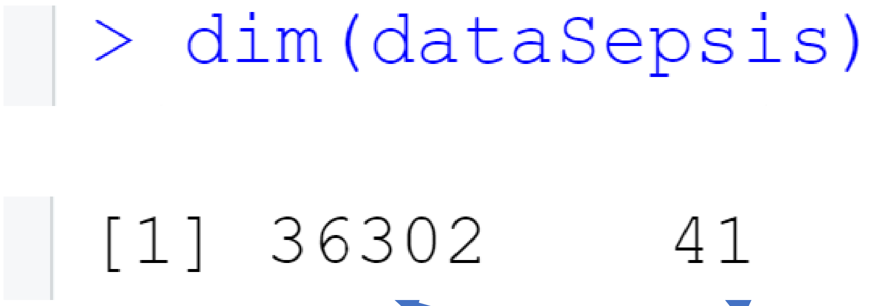
The dim() function provides the contents of the dataset identified within the parentheses. The output consists of the number rows/observations followed by the number of columns/variables.



```
> head(dataSepsis,15)
```

The head() function provides a printout of the top number of rows specified after the comma from the dataset specified in front of the comma, so it takes the form head(datasetname,numberofrows).

Dim() Output



```
> dim(dataSepsis)
```

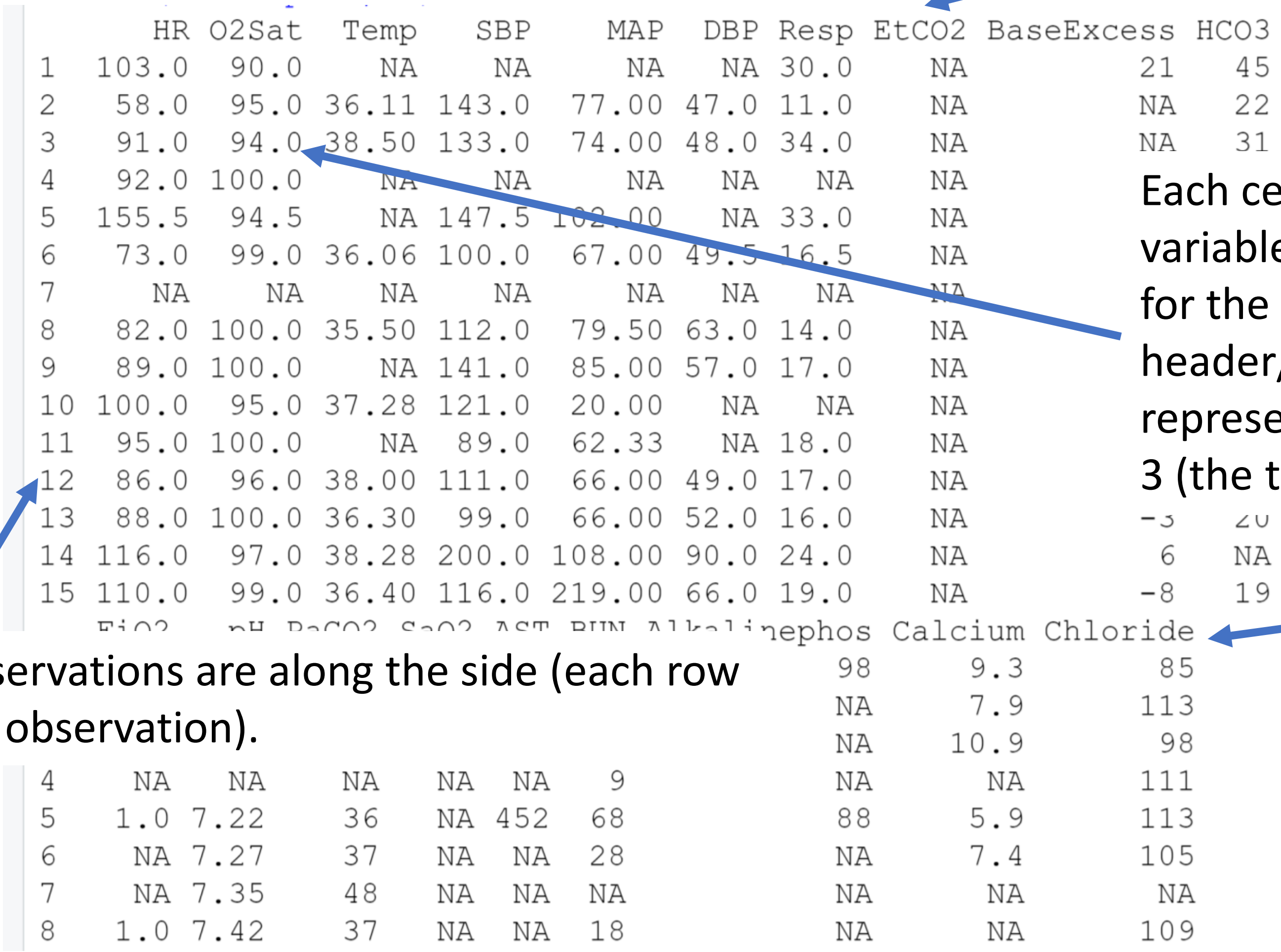
```
[1] 36302     41
```

Code:
dim(datasetname)

Output:
Basic dataset information is here, including the number of observations (rows) and variables (columns).

Head() Output

Variables are across the top (each column is a variable).



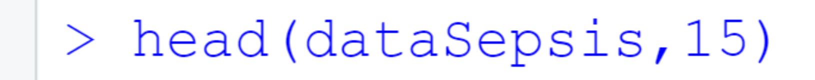
```
        HR O2Sat  Temp   SBP    MAP  DBP Resp EtCO2 BaseExcess HCO3
1  103.0  90.0    NA    NA     NA   NA 30.0    NA         21   45
2   58.0  95.0 36.11 143.0  77.00 47.0 11.0    NA         NA   22
3   91.0  94.0 38.50 133.0  74.00 48.0 34.0    NA         NA   31
4   92.0 100.0    NA    NA     NA   NA   NA    NA
5  155.5  94.5    NA 147.5 102.00   NA 33.0    NA
6   73.0  99.0 36.06 100.0  67.00 49.5 16.5    NA
7     NA    NA    NA    NA     NA   NA   NA    NA
8   82.0 100.0 35.50 112.0  79.50 63.0 14.0    NA
9   89.0 100.0    NA 141.0  85.00 57.0 17.0    NA
10 100.0  95.0 37.28 121.0  20.00   NA   NA    NA
11  95.0 100.0    NA  89.0  62.33   NA 18.0    NA
12  86.0  96.0 38.00 111.0  66.00 49.0 17.0    NA
13  88.0 100.0 36.30  99.0  66.00 52.0 16.0    NA         -3   20
14 116.0  97.0 38.28 200.0 108.00 90.0 24.0    NA          6   NA
15 110.0  99.0 36.40 116.0 219.00 66.0 19.0    NA         -8   19
    FiO2    pH PaCO2 SaO2 AST BUN Alkalinephos Calcium Chloride
                                            98     9.3       85
                                            NA     7.9      113
                                            NA    10.9       98
4     NA    NA    NA   NA  NA   9           NA      NA      111
5    1.0  7.22    36   NA 452  68           88     5.9      113
6     NA  7.27    37   NA  NA  28           NA     7.4      105
7     NA  7.35    48   NA  NA  NA           NA      NA       NA
8    1.0  7.42    37   NA  NA  18           NA      NA      109
```

Each cell represents the value of the variable identified in the column header for the observation identified on the row header/label. This cell, for example, represents an O2Sat of 94 for observation 3 (the third observation).

Columns are continued below

Observations are along the side (each row is a observation).

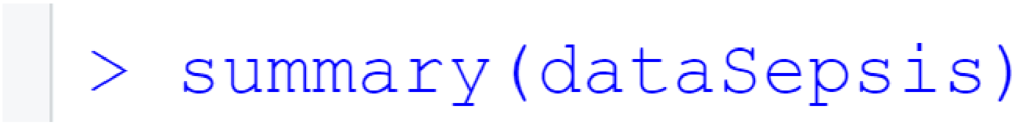
Code:
Head(datasetname,numberofrecords)



```
> head(dataSepsis,15)
```

© Stefanie G. Reay, MS, PhD, Capella University, 2021

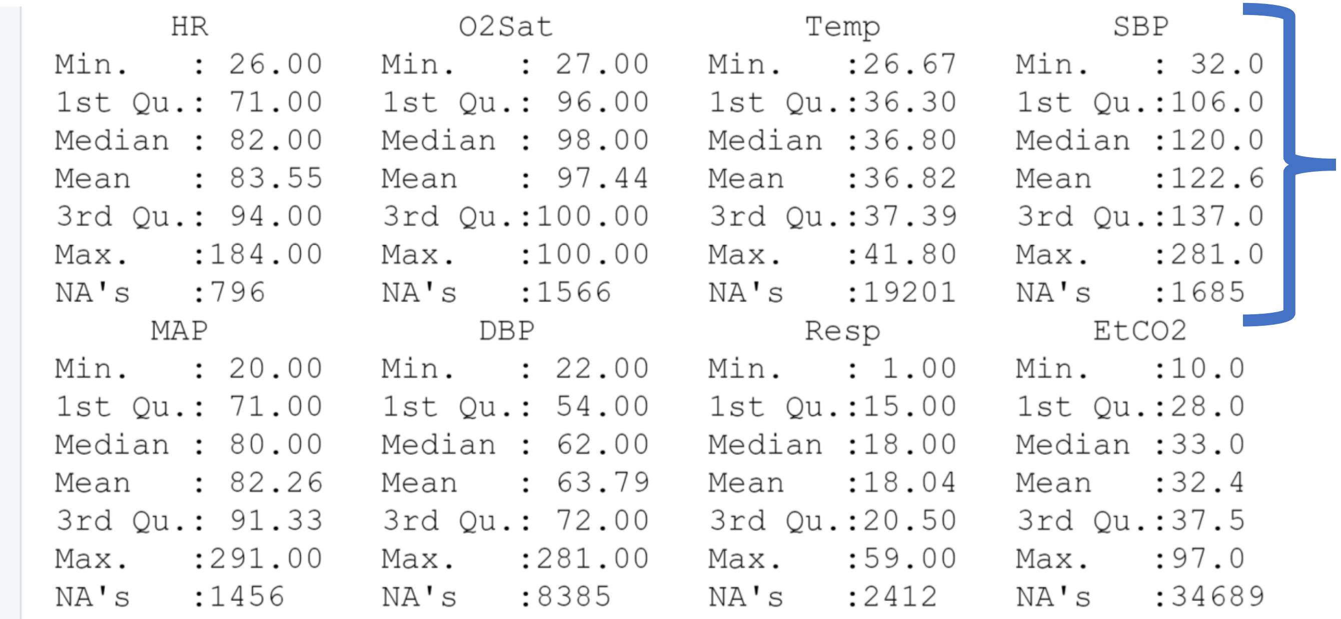
Second: Summarize the variables and check for missing values



```
> summary(dataSepsis)
```

The summary() function outputs several numeric summaries of the variables, including numeric summaries for the numeric variables (minimum, maximum, mean, median, first quartile, and third quartile) and categorical/character variables (length, class, and mode), and the number of null or missing values.

Summary() Output

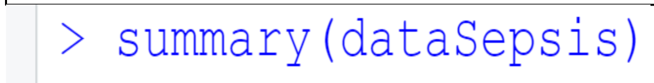


```
      HR              O2Sat              Temp              SBP
 Min.   : 26.00   Min.   : 27.00   Min.   :26.67   Min.   : 32.0
 1st Qu.: 71.00   1st Qu.: 96.00   1st Qu.:36.30   1st Qu.:106.0
 Median : 82.00   Median : 98.00   Median :36.80   Median :120.0
 Mean   : 83.55   Mean   : 97.44   Mean   :36.82   Mean   :122.6
 3rd Qu.: 94.00   3rd Qu.:100.00   3rd Qu.:37.39   3rd Qu.:137.0
 Max.   :184.00   Max.   :100.00   Max.   :41.80   Max.   :281.0
 NA's   :796      NA's   :1566     NA's   :19201   NA's   :1685
      MAP              DBP              Resp             EtCO2
 Min.   : 20.00   Min.   : 22.00   Min.   : 1.00   Min.   :10.0
 1st Qu.: 71.00   1st Qu.: 54.00   1st Qu.:15.00   1st Qu.:28.0
 Median : 80.00   Median : 62.00   Median :18.00   Median :33.0
 Mean   : 82.26   Mean   : 63.79   Mean   :18.04   Mean   :32.4
 3rd Qu.: 91.33   3rd Qu.: 72.00   3rd Qu.:20.50   3rd Qu.:37.5
 Max.   :291.00   Max.   :281.00   Max.   :59.00   Max.   :97.0
 NA's   :1456     NA's   :8385     NA's   :2412    NA's   :34689
```

Basic summary statistics (minimum, maximum, mean, median, first and third quartiles, and the count of null values (i.e. “NA’s”).

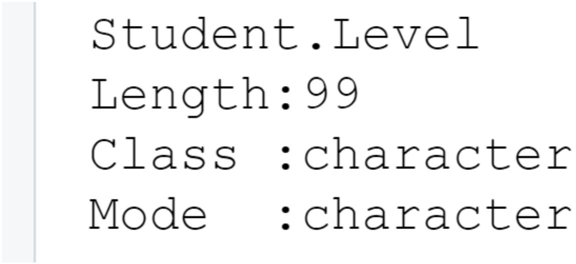
*Note that the output for the sepsis dataset does not have any character variable output, because all of the variables are numeric.

Code:
Summary(datasetname)



```
> summary(dataSepsis)
```

Sample of Summary() Output for a Character Variable



```
 Student.Level
 Length:99
 Class :character
 Mode  :character
```

For a character variable, the summary function output includes the length of the character variable, the class, and the mode.

*Note that the sepsis data did not have any character variables, so this is an example of output from another dataset.

Select Alternative Options

The following package may be installed to assist further with exploratory data analysis:
dplyr
skimr
DataExplorer

Installation instructions for contributed packages are provided in the tutorial for installing and accessing R. To use these packages once they are installed, use the following general template code to call and utlize them:

dplyr:
library(dplyr)
glimpse(datasetname)
*uses the glimps function to display values from the variables within the dataset.

skimr:
library(skimr)
skimr(datasetname)

DataExplorer:
library(DataExplorer)
DataExplorer::create_report(datasetname)
*creates a data exploration report in html format that can be saved as html or printed to PDF. **This is a much more comprehensive data exploration option than using built-in functions within the base R installation.**